串联质谱蛋白质鉴定的关键计算问题

付岩 贺思敏 孙瑞祥 王乐珩

摘要： 蛋白质鉴定是蛋白质组学研究的基础问题，而利用串联质谱搜索蛋白质序列数据库是目前蛋白质鉴定最成功和最常用的方法。蛋白质鉴定软件本质上是一个信息检索系统，具有检索系统的共性，但与文本或多媒体检索相比，又有其非常特殊之处，比如对检索结果进行可靠性评估是蛋白质鉴定必不可少的一步，而这对于其它检索问题往往是不需要的。本文综述蛋白质鉴定搜索引擎中的关键计算问题及其研究进展，包括数据库搜索匹配打分、鉴定结果可靠性统计评估、蛋白质修饰鉴定等，并对我们自己研制的蛋白质鉴定搜索引擎 pFind 做简要介绍。

关键词：生物信息学；蛋白质鉴定；质谱；信息检索；pFind

1 引言

2001 年 2 月，人类基因组计划(Human Genome Project, HGP)组织和美国 Celera 公司分别在《自然（*Nature*）》和《科学（*Science*）》上公布了人类基因组工作草图及初步分析结果。人类基因组测序工作的基本完成，标志着后基因组时代的到来，生命科学的研究在寻找新的生长点。2001 年 4 月在美国成立了以国际合作研究蛋白质组为主要任务的人类蛋白质组组织(Human Proteome Organization, HUPO)，随后各种蛋白质组计划相继展开，包括美国主导的人类血液蛋白质组计划，中国主导的人类肝脏蛋白质组计划，德国主导的人类脑蛋白质组计划等等。同时，针对其它各种生物体的蛋白质组研究也在世界各地广泛开展起来[1]。中国政府将蛋白质科学列为《国家中长期科学与技术发展纲要》四个重大科学计划之一，作为我国 2006 年到 2020 年期间生命科学的研究主题。

“蛋白质组” (Proteome)一词最早是由威金斯（Wilkins）等人于 1994 年首次提出的，用于描述基因组的蛋白质对应物。蛋白质组是指生物细胞、组织或器官在给定时刻和给定条件下表达的蛋白质的全体。顾名思义，蛋白质组学就是对蛋白质组的研究，其最基本的任务就是确定特定有机体内全体蛋白质的状态，包括表达、定量、修饰、突变等方面。蛋白质是由氨基酸分子聚合而成的生物大分子，蛋白质的氨基酸序列唯一确定了蛋白质的身份。大多数的蛋白质在从脱氧核糖核酸（Deoxyribonucleic acid，DNA）经信使核糖核酸（Messenger ribonucleic acid， mRNA）翻译过来之后，还会在特定氨基酸上发生化学修饰，这样才能实现其生物活性。因而，对蛋白质序列的鉴定以及对蛋白质翻译后修饰的刻画对于系统了解蛋白质的结构、功能及进化关系等关键的生物学知识具有十分重要的意义。

生物质谱是目前大规模蛋白质鉴定的主流技术，其优势在于高灵敏度、高通量和高精度等[2]。在典型的自底向上蛋白质组学研究策略中，蛋白质样品被酶解成肽段混合物，后者通过色谱-质谱联用生成串联质谱。从串联质谱重构出肽段序列，是蛋白质鉴定的核心计算问题。目前，最成功和最常用的解决方法是用串联质谱搜索蛋白质序列数据库，将数据库中的序列做理论酶切和理论碎裂，然后将预测的谱图跟实验谱图匹配，从而鉴定肽序列，进而鉴定整个蛋白。基于蛋白质序列库搜索的蛋白质鉴定，实际上是一个检索系统，其核心计算问题是谱图匹配的肽打分算法。同时，为了得到正确的鉴定结果，蛋白质鉴定系统还必须对检

索结果的可靠性进行统计评估。蛋白质的修饰给蛋白质鉴定检索系统的速度和精度都带来了更大挑战。本文下面主要就从这几方面综述蛋白质鉴定中的关键计算问题及目前的解决策略，在这之前先简要介绍相关的生化背景。

2　生化背景知识

1.1 蛋白质和肽

蛋白质是一切生命的物质基础，广泛存在于各种生物组织细胞中，是生物细胞最重要的组成物质。蛋白质是一类重要的生物大分子，是生物体内结构和功能的主要载体。人体中蛋白质多达 10 万种以上，结构和功能千差万别。但是，所有蛋白质都是由叫做氨基酸的分子连接而成的。氨基酸分子的通式如图 1 所示。氨基酸是由α-碳原子，以及与其相连的羧基(−COOH)、氨基(−NH)、氢原子(H)和侧链基团 R 构成的。不同的氨基酸具有不同的侧链基团。一个氨基酸的羧基可以与另一个氨基酸的氨基缩合脱水形成酰胺键(称为肽键)而连接起来，如图 2 所示。多个氨基酸以肽键顺序相连，形成链状分子，称为肽，通常称氨基端为 N 端，羧基端为 C 端，如图 3 所示。由两个氨基酸构成的肽称为二肽，由 2 到 10 个氨基酸构成的肽成为寡肽，由 10 个以上氨基酸构成的肽成为多肽。分子重量在 10K Da 以上的多肽称为蛋白质。

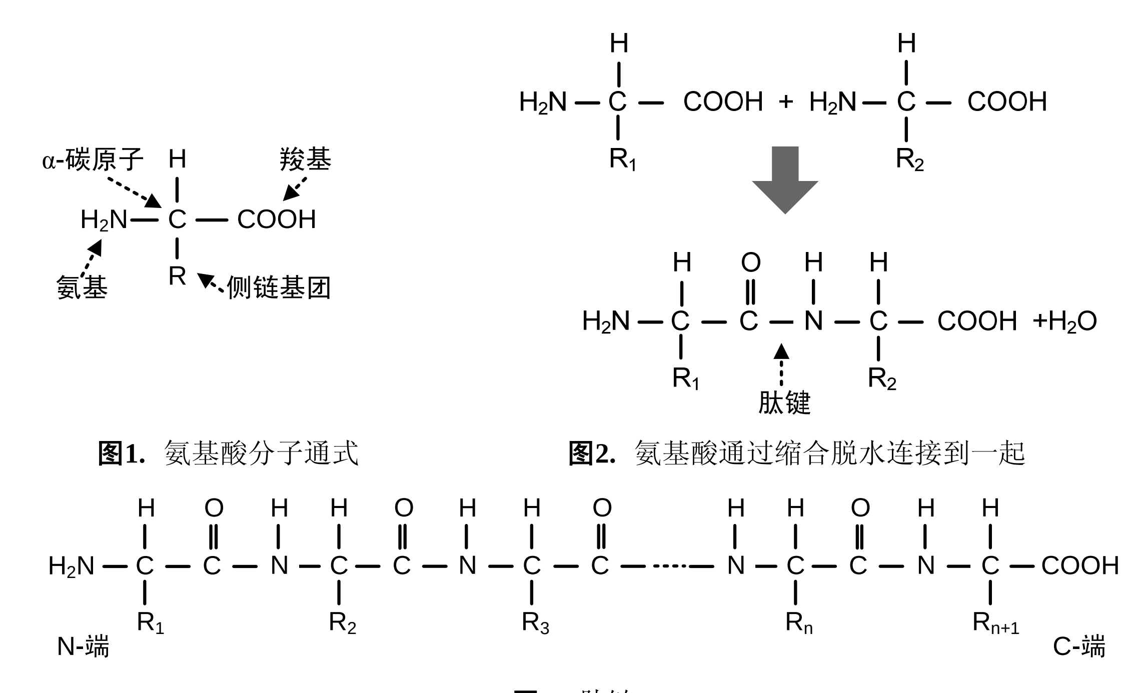


图1. 氨基酸分子通式

图2. 氨基酸通过缩合脱水连接到一起

图3. 肽链

绝大多数的蛋白质是由常见的 20 种氨基酸组成的。有少数蛋白质包含几百种不常见的氨基酸以及非肽链结构的其它组成成分(称为配基或辅基)。蛋白质的氨基酸序列称为蛋白质的一级结构。蛋白质可以通过折叠等形成二级和三级等空间结构。蛋白质的一级结构，即蛋白质的氨基酸序列(简称蛋白质序列)，唯一确定了蛋白质的身份。本文所述的蛋白质鉴定问题，就是指对蛋白质序列进行鉴定。

1.2 生物质谱技术

最初，蛋白质序列鉴定主要采用手工的埃德曼降解-环甲基化方法，效率很低。质谱技术(Mass Spectrometry，MS)的发展为蛋白质序列鉴定开辟了新的途径[3, 4]。

质谱技术的基本原理其实并不复杂。在质谱分析中，待分析的物质粒子首先被离子化，然后再通过适当的电磁场。由于不同质量电荷比的离子对电磁场的反应不同，因而可按照运动轨迹和时间等进行分离和检测。离子的强度同时也被检测和记录。从而得到以离子质量电荷比为横坐标，以离子强度为纵坐标的质谱数据。质谱仪由进样系统、离子源、分离系统和检测系统四大部分构成，每个部分都有多种实现方式。图 4 简单描绘了质谱仪的构成。

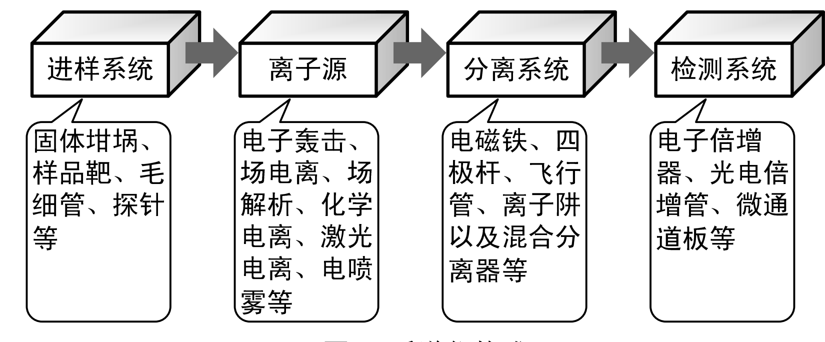


图4. 质谱仪构成

质谱技术的历史可追溯到 19 世纪末。1899 年汤姆逊(Joseph John Thomson)发明了第一台抛物线质谱装置。随着技术的改进，20 世纪 50 年代后期质谱仪广泛地应用于无机化合物和有机化合物的测定。二十世纪 50-80 年代，质谱技术进入繁荣时期。到了 20 世纪 80-90 年代，质谱技术得到革命性的发展，主要是液相-质谱联用技术以及电喷雾离子化(Electrospray ionization, ESI)和基质辅助激光解吸附离子化(Matrix Assisted Laser Desorption/Ionization, MALDI)两种软电离技术的发展。20 世纪 80 年代，芬恩（John Fenn）和田中耕一（Koichi Tanaka）分别发明了能够用于生物大分子分析的 ESI 质谱技术和 MALDI 质谱技术，他们因此与维特里希（Kurt Wüthrich）共同获得了 2002 年的诺贝尔化学奖。20 世纪 90 年代开始，质谱技术在生命科学领域得到了深入应用。

在用于蛋白质分析的质谱技术中，蛋白质样品首先被选定的蛋白酶水解，形成多肽。不同质量电荷比的多肽离子被质谱仪分离、检测出来，得到一级质谱。这些肽离子可以进一步被打碎，形成碎片离子。碎片离子被分离和检测便得到串联质谱。用质谱鉴定蛋白质的方法因此分为两大类。

第一类是基于一级质谱的，称作肽质量指纹作图。这类方法搜索已知蛋白质数据库，用指定的酶对蛋白质进行模拟水解，得到理论一级质谱。理论一级质谱与实验一级质谱进行比较，结果按照匹配的程度排序。这样的系统有 MOWSE[5]，Mascot[6]，ProFound[7]，PeptIdent[8]，MS-Fit[9]等等。肽质量指纹作图适用于蛋白质样品包含一种蛋白质或简单的混合物的情况。其缺点是由于蛋白质混合物和污染物、部分酶解、残基修饰（所谓氨基酸残基是指去掉一个水分子的氨基酸）、质量精度等因素的影响，往往误差较大，导致搜索结果错误。

第二类是基于串联质谱的。这类方法首先利用串联质谱技术(Tandem Mass Spectrometry, MS/MS)准确测定肽的氨基酸序列，再通过肽序列鉴定蛋白质的序列。所以，这种方法可以用于鉴定复杂的蛋白质混合物或者验证肽质量指纹搜索的结果，是目前最常用最有效的主流方法，下面加以详细介绍。

在典型的液相色谱-串联质谱联用实验中，蛋白质样品首先被蛋白酶水解得到多肽混合物，然后通过液相色谱分离并被离子化。在质谱仪中，具有特定质量电荷比的肽离子被选择过滤后，在某种能量轰击，比如碰撞诱导的裂解(Collision-Induced Dissociation，CID)[10]或电子转运裂解(Electron Transfer Dissociation，ETD)[11]，作用下碎裂。在碎裂过程中，三种肽键断裂能够生成主要六个系列的碎片，即 N-端的 a，b，c 碎片和 C-端的 x，y，z 碎片，如图 5 所示。碎片可能丢失一个中性的水或者氨分子[12]，保留了母离子电荷的碎片离子、没有碎

裂的母离子、污染物、以及碎裂产生的其它类型的离子被检测出来。在低能量裂解时，一个肽离子上一般只有一处发生肽键断裂。产生的离子类型主要是 a，b 和 y 型碎片离子。通过测量具有不同质量电荷比的离子的强度形成串联质谱中的谱峰。图 6 是串联质谱的一个例子。

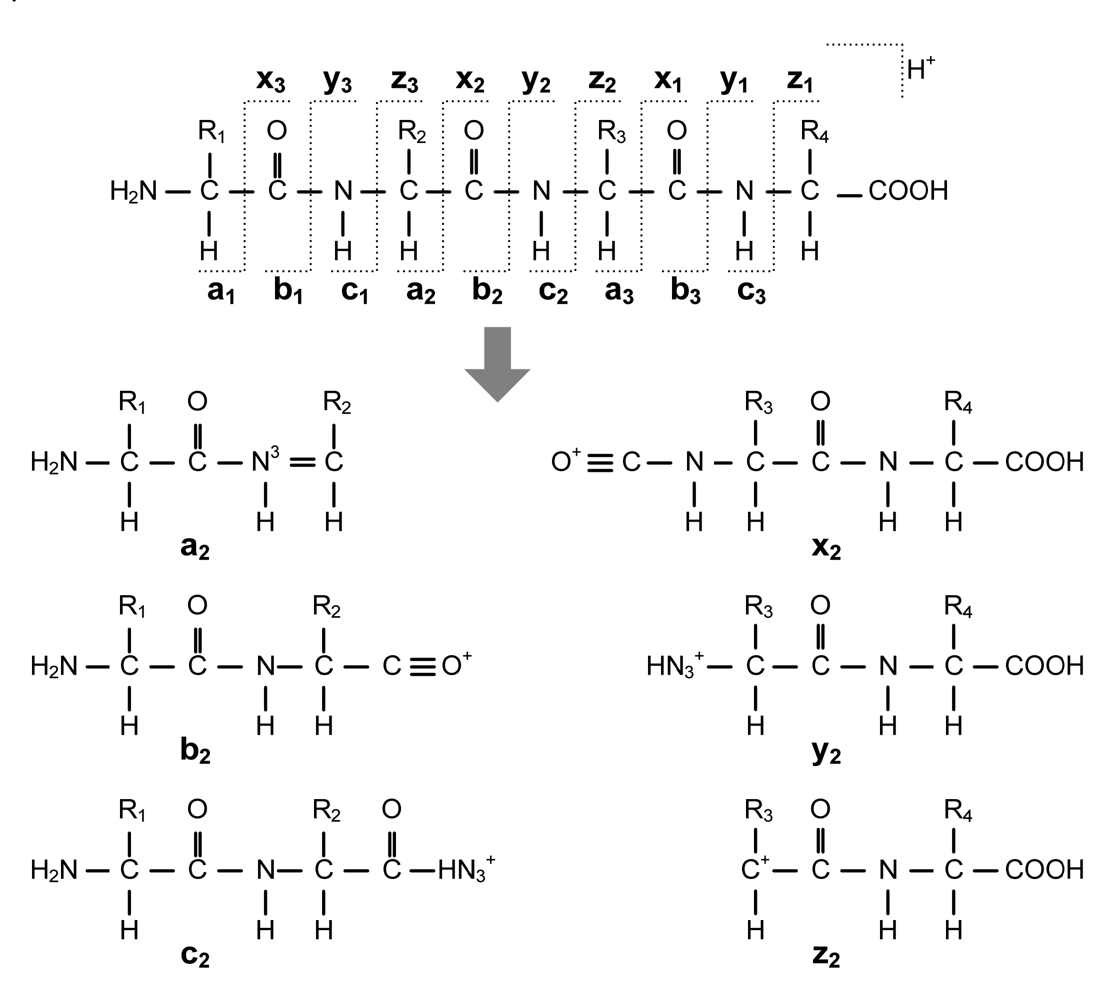


图5. CID 作用下肽离子碎裂形成的碎片离子

为了鉴定蛋白质，从串联质谱鉴定肽的氨基酸序列是中心问题。从串联质谱鉴定肽序列的计算方法有三种。最常用的是数据库搜索方法，如文献[6, 9, 13-18]。在这种方法中，数据库中的蛋白质序列被理论水解和碎裂，生成理论串联质谱。把理论质谱与实验质谱相比较，从而找到生成实验质谱的肽序列。本文就是针对这种方法进行介绍。

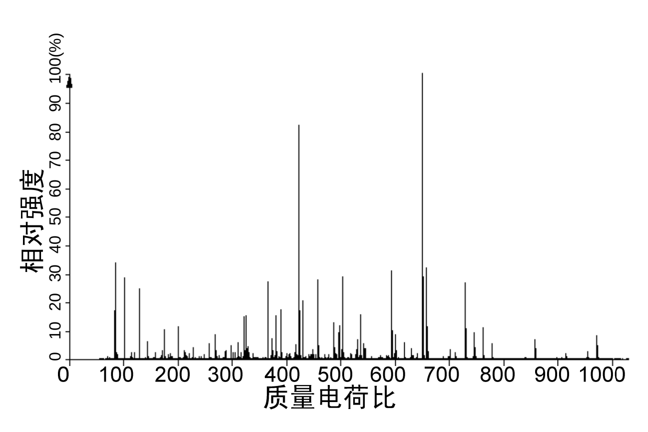


图6. 串联质谱示例

第二种是从头(de novo)测序方法，它通过直接解释串联质谱数据来进行肽序列鉴定，而不是与数据库中的序列进行比较，比如文献[19-26]等。当数据库中没有目标序列时，搜索数据库的方法就无能为力了，所以不得不使用从头测序的方法。但是该方法的难点在于要求质谱数据有较高质量，肽断裂情况良好，所以目前尚未得到广泛的实际应用。不过从头测序方法即使不能完全测出肽序列，也能

提供重要的肽序列标签(几个氨基酸组成的短肽片段)，供数据库搜索参考。

第三种方法是序列标签查询的方法[27-31]，首先从串联质谱中人工或自动地获得肽序列的片段信息，然后利用这些部分序列信息查询数据库，得到肽的全序列。这种方法是前两种方法的结合，近年来受到越来越多的关注。

关于基于串联质谱的肽鉴定，最近有一些比较详尽的综述文献[32-35]。

3 数据库检索打分算法

在利用串联质谱鉴定蛋白质的方法中，蛋白质鉴定问题归约为更基本的肽鉴定问题。而数据库搜索方法是目前普遍采用的肽鉴定方法。给定实验串联质谱，对数据库中的候选肽进行匹配打分是肽鉴定算法的核心。对肽打分鉴定结果的评价，即识别出正确鉴定的肽序列，也是必不可少的一步。

所谓“肽打分”是指：给定实验串联质谱，对候选肽产生该质谱的可能程度做出评分，从而对所有候选肽进行排序。用信息检索的语言描述，这里的串联质谱即是输入的查询，候选肽即是数据库中保存的对象，而肽打分函数实际上就是检索函数或称排位函数。肽打分函数的功能就是对候选肽进行排位，把最可能产生实验质谱的肽序列排在首位。可以把肽打分函数按构造方式分为三类。第一类基于谱向量点积；第二类基于概率；第三类基于机器学习，或者说基于模式分类。

3.1 基于谱向量点积的肽打分算法

在基于谱向量点积(spectral dot product，SDP)的肽打分算法中，是把理论质谱和实验质谱重叠的程度作为候选肽的分值，而这种重叠可以用向量间的点积运算描述。在 SDP 中，理论和实验质谱分别被表示为 N 维向量 $c=\left[c_1,c_2\cdots,c_N\right]$ 和 $t=\left[t_1,t_2\cdots,t_N\right]$ [36]。其中，N 是所使用的不同质量值的数量，c_i 和 t_i 可以取 0/1 值，也可以取串联质谱中第 i 个质量值的离子强度。实验和理论串联质谱间的 SDP 定义为：

$$SDP=\boldsymbol{c}\cdot\boldsymbol{t}=\sum_{i=1}^{N}c_i t_i$$

如果两个谱向量是相同的，那么它们应该是平行的。而向量的点积恰好反映了它们平行的程度，因此可作为肽匹配的分值。

在文献[36]中，基于 SDP 的谱差角被用作质谱的相似性度量。文献[37]利用这种度量识别由相同肽序列产生的质谱。早期使用的“共有峰计数”(Shared Peaks Count，SPC)打分方法就是谱向量点积的最简单形式。所谓 SPC 是指理论和实验质谱之间匹配的碎片离子的数目。所以，SPC 对应于 SDP 中 c_i 和 t_i 取 0/1 值的情况。Sonar MS/MS [16, 38]软件是使用 SDP 作为肽打分函数的典型代表，它将质谱表示成向量形式并直接计算谱向量的点积作为分数。

目前使用最广泛的商业肽鉴定软件之一 SEQUEST[14]是利用信号间的交叉相关分析来比较质谱的，而其中的交叉相关运算实际上也是基于谱向量点积的。首先按一定规则对匹配的氨基酸序列预测其质谱，再对实验质谱做适当处理，以使两个质谱之间交叉-相关分析能够反映出碎片离子的相似度。作为离散信号的实验谱 $x(t)$ 和理论谱 $y(t)$ 之间的交叉-相关如下计算：

$$R_\tau = \sum_{i=0}^{n-1} x[i] y[i+\tau] \tau$$

其中，τ 是两个信号间的位移值。相关函数实际上是测量了两个信号间的相似度。如果两个信号是相同的，则相关函数在 $\tau = 0$ 处取最大值。SEQUEST 打分公式定义为：

$$X_{corr} = R_\tau(\tau = 0) - \frac{1}{149} \sum_{-75<k<75} R_\tau(\tau = k)$$

可见，X_{corr} 分值在实际上就是 SDP 再减去一系列位移的 SDP 的均值。

3.2 基于概率的肽打分算法

另一类肽打分算法是基于概率的，如 Mascot [6]，SCOPE[13]，ProbID[18]和 PepSearch[39]，以及文献[40]等等。Mascot 是除了 SEQUEST 之外，另一个广泛采用的商业蛋白质鉴定软件。但是在关于 Mascot 的文献里，并没有具体给出 Mascot 采用的肽打分算法。总体来说，Mascot 试图计算实验串联质谱由候选肽随机生成的概率 p，而候选肽分数为 $-\log(p)$。Mascot 的概率打分算法综合考虑了肽长度的分布、酶切位点遗漏概率、质量误差分布以及离子强度等因素。

SCOPE 是 Celera 公司设计的打分算法。它利用贝叶斯模型进行打分，对于给定的质谱求每个序列的后验概率。SCOPE 通过用两步随机过程模拟串联质谱生成的过程：1)根据概率分布生成肽的碎片；2)根据仪器测量误差，从碎片生成质谱。

ProbID 试图计算实验串联质谱由候选肽随机生成的贝叶斯后验概率。但是 ProbID 计算的概率不能算作真正意义上的概率，而只不过是若干因素的简单乘积。其中包括亚胺离子的出现情况、肽序列酶切点是否满足酶的特异性、匹配和不匹配的谱峰以及连续和互补离子的匹配情况等等。

SCOPE 和 ProbID 虽然在不同层次上建立了打分的概率模型，但是它们的共同点是用于计算的条件概率，如不同离子出现的概率、误差分布的概率、离子强度的概率分布等都是根据专家经验指定或假定的，因而是不准确的。

哈维里欧（Havilio）等人[40]和丹西克（Dancik）等人[20]试图从质谱数据中学习这些概率。文献[20]从数据中学习检测到某种碎片类型的概率，而不是先验假定的。文献[40]中的方法是对文献[20]中算法的推广，设计了一系列打分函数，可以包含有关肽碎裂的各种各样的实验观测和先验理论知识，考虑了强度间的相关性，碎片类型、碎片质量以及碎片质量与肽质量之比，重要的离子类型，如同位素和多电荷碎片等，这些经常被串联质谱分析软件忽略。学习参数的过程是自动的，即把质谱的质量轴划分成等宽的小片断，对所有落在片断上的离子，计算其观测强度的概率。如果假设碎片独立，则把所有概率相乘。如果有相关碎片对，则计算相关离子对的联合概率。这种做法的缺点在于质谱数据没有经过标注，只是粗略地把所有与某个谱峰都匹配的离子进行统计。所以，其统计结果必然是不够准确的。其它挖掘质谱数据的工作中也存在相同的问题[41]。

上面几种基于概率的肽打分算法是对肽碎裂产生质谱的过程进行概率建模。另一类基于概率的肽打分算法则不对肽碎裂的过程进行建模，而是针对预测离子与谱峰的匹配进行概率建模。比如，萨迪戈夫（Sadygov）和耶茨（Yates）采用超几何分布[42]，弗里德曼（Fridman）等人则采用一种更复杂形式的超几何分布[43]，而吉尔（Geer）等人采用泊松分布[44]。这类基于概率的肽打分算法的优点是能够给出候选肽与实验质谱正确或随机匹配的概率，但是对

理论质谱预测和谱峰强度信息的利用不够。

3.3 基于机器学习的肽打分方法

肽鉴定本质上可以看作将候选肽分为“正确”和“不正确”的两类分类问题。在基于机器学习的肽打分函数中，候选肽与实验质谱间的多种匹配信息被表示成特征向量的形式，然后利用机器学习方法从序列已知的质谱训练数据中学习出一个打分函数。虽然信息检索中的检索函数机器学习方法早就存在了，但是直到最近才被应用到肽鉴定问题中来。多种机器学习算法被应用在 SEQUEST 软件搜索结果的分类上，如支持向量机（support vector machine，SVM）[45]，神经网络[46, 47]，逻辑回归（Logistic regression）[48]，以及 boosting 算法[1]和随机森林（random forest）[2]等集成方法[49]。使用机器学习方法进行肽鉴定打分的好处是可以综合利用很多种匹配指标，将每种指标作为模式的一个维度。如何把这些指标融合成一个肽打分函数，完全成为机器学习方法的任务，不需要用户的介入。实际上，如何把众多的匹配指标综合成一个肽打分函数，一直是肽打分设计的难点之一。鉴于机器学习方法的灵活性，后来有研究者开始利用机器学习方法直接构造独立的肽打分函数，而不是仅仅满足于对已有肽鉴定软件搜索结果的后处理[50, 51]。实际上，基于数据库搜索的肽鉴定本身就是一个具体的信息检索问题，因此用机器学习方法直接构造独立的肽打分函数本质上就是信息检索中的检索函数或者排位函数（Ranking function）学习问题。当然，检索函数可以是判别性的，也可以只有排序功能。

4 鉴定结果可靠性的统计评估

对于每个质谱，在数据库搜索之后，尽管总会有一个得分最高的候选肽，但是这个候选肽不一定就是正确的。造成这种结果的可能原因有很多，比如：肽打分算法是不完美的，总会有犯错误的情况，没能把正确的肽序列排第一名；搜索的蛋白质序列库是不完全的，不包含目标肽序列；输入的质谱数据完全由噪音产生，不包含有效信息；目标肽发生了未预料到的修饰，或由不正常的酶切产生，等等。所以，在肽打分后，应该确定获得最高分的肽是否是正确的答案。也就是说，要对肽打分结果的可靠性进行评估，找出正确的肽鉴定结果。

对肽鉴定结果可靠性的评估早期使用的是经验阈值法。顾名思义，经验阈值法就是根据经验对原始打分结果施加一个阈值，得分在阈值以上的肽才被认定为正确的鉴定结果。典型的例子是 SEQUEST 软件[14]。SEQUEST 输出的两个主要分值是 *Xcorr* 和 *DeltCn* 。过去，这两个分值被广泛用于 SEQUEST 肽打分结果的过滤。比如，一种常用的过滤准则是要求 *DeltCn* 大于 0.1，同时，对于带一个、两个和三个电荷的肽， *Xcorr* 要分别大于 1.9、2.2 和 3.75[52]。针对 SEQUEST，也有在后处理步骤中计算更好的分值来进行过滤的方法，如 RScore[53]。在基于概率的肽打分算法中，虽然计算的目标是真实匹配或随机匹配的概率，但是前面已经指出，由于各种各样的原因，这样的概率客观上是无法准确计算的。因此，基于概率的肽打分方法通常仍需要指定一个阈值，或者使用附加的评价方法。经验阈值法的好处是简单直观，但缺点也很明显，那就是阈值的指定只凭经验，缺少理论上的依据。当数据库规模增大时，错误候选肽的最高分也会水涨船高。并且，根据阈值过滤出的结果，其可靠性没有定量的估计。使用经验阈值是一种武断的做法，实际上，无论肽鉴定结果的得分有多高，都带有或多或少的不确定性。为了有效估计鉴定结果的可靠性，必须利用统计手段。目前，使用最多的鉴定结果可靠性统计度量指标是针对单谱鉴定的期望值和针对多谱鉴定的假发

[1] 基本思想是将多个能力较弱的分类器迭加，得到一个更强的分类器
[2] 一个包含多个决策树的分类器

现率[3](False discovery rate , FDR)。

4.1 期望值方法

期望值（通常缩写为 E-value）指一个随机变量的平均值。在生物信息学领域的序列比对问题中[54-56]首先成功应用了期望值方法，序列比对程序 BLAST 最初使用了期望值来度量序列比对得分随机发生的可能性(http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html)。在肽鉴定中，给定一张谱图和 n 条随机候选肽，我们感兴趣的随机变量是得分至少为 x 的错误候选肽的数目。在含义上，分值 x 的期望值就是能以随机的方式得到等于或超过分值 x 的候选肽的数目的期望。比如，假设在一次数据库搜索中一个候选肽得分的期望值为 10^{-5}，这意味着，平均要做十万次这样的搜索，才能随机地得到等于或大于这个分值的得分。所以，从理论上讲，如果一个候选肽对应的期望值大于 1，就可以排除这个候选肽了，因为即使是完全随机的情况，平均也会有一个候选肽得到相同或更高的分值。

随机分值的概率分布决定了期望值的计算方法。依据概率分布来源的不同，期望值的计算方法可分为三类，即经验拟合法、理论建模法和穷举计算法。经验拟合法是通过拟合实际的分值分布数据估计概率分布，理论建模法是通过假定某种随机概率模型从理论上推导分值的随机分布，穷举计算法是通过穷举所有可能候选肽得到分值的真实分布。通过经验拟合法计算期望值的搜索引擎有 Sonar[38]，X!Tandem[57]，pFind[58-60]，及 RAId_DbS[61]等，通过理论建模法计算期望值的搜索引擎有 Mascot[6]和 OMSSA[44]，穷举计算法则是最近由金（Kim）等人提出的[62]。三种方法各有优缺点，经验拟合法适用于任意类型的打分函数，但要求必须有足够的候选肽规模以拟合分值分布以及恰当的分布形式假设；理论建模法对于任意给定的一条肽序列都可以计算其期望值，但是只适用于基于概率的打分函数，准确性取决于概率模型的准确性；穷举计算法能够直接计算出分值的真实分布，但是只适用于加和性的打分函数，并且计算复杂度较高。经验拟合法是目前最常用和最成功的期望值计算方法，下面予以简单介绍。

在肽鉴定数据库搜索中，给定一张谱图，至多只能有一个候选肽是正确的，所以可以认为几乎所有的候选肽都是错误的。经验的期望值计算方法就是直接把一次数据库搜索产生的所有候选肽分值用于随机分值分布的拟合[38]。假设错误候选肽的分值满足一定形式的概率分布，并从经验数据中估计出参数，就可以推测出得分不小于 x 的概率 P_x，进一步乘上参与打分的候选肽数目，就得到期望值。

文献[38]假设随机分值 x 服从极值分布。根据这个假设，P_x 与 x 在高分值区域有近似的对数线性关系，即

$$\log\left(P_x\right) = c_1 \log x + c_2$$

其中，系数 c_1 和 c_2 的值可以从一次搜索中候选肽分值分布的高端部分估计。当 P_x 估计出来之后，对任意候选肽的分值，就可以计算其期望值了。X! TANDEM 软件就采用了这个方法来计算期望值[57]。

但是，并不是所有的打分算法都符合上述的分布假设。实际上，关于上述期望值计算方法在肽鉴定问题中的适用性以及如何拟合分值分布最近引起了一些讨论[63, 64]。虽然期望值有明确的定义，但是各个软件计算出的期望值并没有严格的绝对意义，因为在计算中总要做一些假设和近似。实际上，各个软件在计算期望值时采用的假设不同，实现方法也有差别，

[3] 亦有译作“错误发现率”或“假阳性率”

造成输出的期望值并不能直接互相比较[65]。但是，这并不说明期望值失去了其优越性。与原始打分相比，期望值考虑了分值分布和数据库的规模，因而可以看作归一化后的相对分值。我们要做的是让期望值的估计值尽可能地接近真实值。

4.2 假发现率方法

上面介绍的期望值实现了对单个肽鉴定结果的可靠性评估。但在蛋白质组实验中，一次性鉴定的谱图往往不是一个而是成千上万，对于用给定的肽打分（或期望值）阈值过滤得到的大量肽鉴定结果，需要从整体上评估其可靠性。目前，针对肽鉴定结果群体可靠性的评估通常采用计算假发现率[66]的方法。FDR 的计算可分为两大类，一类是拟合某种分值分布模型来估计后验错误概率；另一类是通过引入诱饵序列库作为对照。从机器学习的角度讲，前者是无监督的，后者是有监督的。

PeptideProphet 是最具代表性和最成功的基于模型的假发现率估计方法[67]。每次蛋白质组实验会产生大量的串联质谱，通过数据库搜索，每个质谱都会被分配到一个得分最高的候选肽。PeptideProphet 方法就是基于对这些候选肽最高分值分布的分析。在 PeptideProphet 方法中，SEQUEST 给出的几种分值首先被线性合并为一个判别分值，并假设错误匹配的判别分值服从伽玛分布，而正确匹配的判别分值服从高斯分布。针对每次具体实验，PeptideProphet 使用期望最大化(EM)算法对判别分值的分布进行参数估计，从而找到能在最大程度上区分正确和错误匹配的分值阈值，同时对错误率做出估计。

目前使用更普遍的假发现率计算方法是基于诱饵（decoy）序列库搜索的方法。所谓诱饵序列是指一定不包含目标蛋白的序列，搜索这样的序列得到的结果一定是错误的结果，因而可作为阴性样本来估计假发现率。诱饵序列通常是目标序列的反转，或者随机生成的序列，或者是根据某种概率模型生成的序列。诱饵序列的特点是，必须不包含目标序列，同时又具有目标序列的“特征”。这样，估计的假发现率才准确。目前，利用反转库估计假阳性率的方法简单而实用，已经被蛋白质学界广泛采用，成为蛋白质组数据假阳性分析的一种标准[68, 69]。利用反转库方法估计假发现率的步骤如下：

1. 将包含目标蛋白的数据库中的序列反转，得到反向序列；并将反向序列与正向序列合并，形成所谓的目标-诱饵(target-decoy)数据库；
2. 用任意肽鉴定软件搜索目标-诱饵数据库，对一次蛋白质组实验产生的所有质谱进行鉴定；
3. 采用任意单谱评价方法对肽打分结果进行过滤，得到阳性肽鉴定结果；
4. 估计阳性肽鉴定结果的假发现率：令 N_f 表示肽序列来自正向蛋白序列的阳性肽鉴定数目，N_r 表示肽序列来自反向蛋白序列的阳性肽鉴定数目(如果肽打分和过滤算法都是有效的，应有 $N_r < N_f$，实际上，一般是 $N_r << N_f$)，则肽鉴定结果的假发现率为：

$$FDR = \frac{2N_r}{N_r + N_f} \times 100\%$$

上面的假发现率计算公式是基于这样的假设：正确鉴定中肽序列一定来自正向蛋白序列，而错误鉴定的肽序列来自正向蛋白序列或反向蛋白序列的可能性是一样的。因为正向序列和反向序列的长度是一样的。所以，可以认为正向序列肽鉴定结果中包含了数目与反向序列肽鉴定数目相同的假阳性鉴定结果。

5 蛋白质修饰鉴定

蛋白质在从信使核糖核酸翻译形成后，可能会在某些氨基酸上增加某种功能团，或增加了其它的蛋白质或肽，或改变了氨基酸的化学性质或结构。这一过程被称为发生了化学修饰。由于该过程发生在翻译过程之后，因此被称为蛋白质的翻译后修饰（PTM，Post-Translational Modification）。翻译后修饰能够改变氨基酸的化学性质，引起蛋白质结构的改变，调控蛋白质的活性和功能。翻译后修饰在生物体内的存在非常普遍，绝大多数的蛋白质都会含有一个或多个翻译后修饰。研究翻译后修饰对于阐明蛋白质的功能，解释重大疾病的发生机理等具有十分重要的意义[70, 71]。对人类蛋白质组的研究表明，对于较高（>1%）表达水平的胰蛋白酶酶切肽段，平均每个氨基酸都几乎有一种修饰形式[72]。除了体内发生的修饰，在样品处理中也不可避免地会引入很多种修饰[73]。蛋白质修饰的种类繁多，截止至 2009 年 6 月 24 日，Unimod 修饰数据库中已有 590 条记录。基于质谱技术的蛋白质组学为大规模翻译后修饰研究提供了有效的分析手段[74-76]。目前，利用串联质谱数据鉴定发生修饰的蛋白质已经成为蛋白质组学研究的核心和前沿问题之一。

为了鉴定发生翻译后修饰的蛋白质，一种常见的基于串联质谱的鉴定方法是在数据库搜索时指定一些可变修饰类型，然后在生成候选肽时同时考虑发生和不发生指定修饰的情况，当候选肽中有多个可能的修饰位点时考虑所有可能的组合。这种方法考虑到了蛋白质翻译后修饰的动态性（相同的氨基酸位点可能发生某种修饰，也可能不发生），但由于天然存在或人工引入的修饰类型有几百种。所以，在数据库搜索时考虑过多的修饰类型是不现实的。这会导致搜索空间组合爆炸，大大降低数据库搜索的速度，同时导致假阳性搜索结果增多。现有技术中的相应搜索引擎，如 SEQUEST 和 Mascot，容许指定的可变修饰类型一般不超过 10 种，这显然不能满足实际需要。在一般情况下，实验人员对蛋白质样品中存在的修饰类型知之甚少，主要依靠经验猜测。大多数时候，蛋氨酸上的氧化修饰是数据库搜索时唯一指定的可变修饰。这样就可能会遗漏样品中存在的其它修饰类型。同时，很多由修饰肽产生的质谱数据得不到解析。这种指定若干种修饰类型的做法被称为限制性修饰鉴定，具有盲目性、搜索空间组合爆炸、不能发现新类型修饰等严重问题。

蛋白质序列数据库太大和可变修饰类型数目太多，共同导致了候选肽空间组合爆炸的问题。如果把搜索限于较小的蛋白质数据库，则可以多考虑一些可变修饰类型。一种常用的办法是二次精细数据库搜索[77-80]。在第一次搜索时，搜索整个蛋白质数据库，但仅考虑最少的可变修饰类型以及最严格的酶切方式。在第二次搜索中，用在第一次搜索中鉴定出的肽所在的蛋白质组成一个小的数据库，在其上进行精细搜索，考虑更多的可变修饰类型，以及宽松的酶切方式和序列突变等。这种方法最早在 MASCOT 软件中使用[77] 。其基本假设是实验样品中含有的每个蛋白质至少有一个肽段可在第一次搜索中被鉴定出来[78]。这种先粗糙搜索，再精细搜索的策略，可以大大提高搜索的速度和可变修饰类型的种类。同时，因为考虑了更多的修饰类型以及酶切方式和序列突变，可以鉴定出更多的肽及其变体。可谓一举双得。但是，如果上述假设不满足，则可能漏掉一些肽和蛋白质及其变体。而且，这种二次搜索仍需要用户指定一个可变修饰类型列表，仍属于限制性的修饰鉴定。所以，无法检测出列表之外的和未知的修饰类型。

为了使数据库搜索方法能够应用于未知的或预料之外的修饰类型鉴定，最直接的办法就是放开肽-谱匹配的肽质量限制，让正确的候选肽序列进入搜索空间，跟实验谱图进行匹配操作。这样做无疑大大增加了计算量，这个问题稍后再讨论。更重要的是，如何把含有修饰的实验谱图跟没有修饰的候选肽序列进行匹配，使得正确的候选肽序列能够被发现，并确定

修饰质量和位点。MS-Alignment 是此类方法中最早和最为著名的[17, 81, 82]。MS-Alignment 以一种类似基因组学中序列比对的方式，将理论质谱与实验质谱相比对，允许任意修饰的出现。

但是，MS-alignment 算法有几个方面的不足： 1）寻找实验谱跟肽序列的最优匹配的计算复杂度很高，数据分析速度非常慢，实际应用中只能针对非常少量的蛋白序列进行搜索；2）为了使用动态规划算法比对理论谱和实验谱，不得不使用简单形式的打分函数，降低了谱图比对的准确性；3）搜索结果的可靠性低，陈等人[83]最近的一项研究显示，MS-alignment 算法严重低估了结果的假发现率；

由相同序列肽的修饰和非修饰形式分别产生的谱图即为一种典型的相关谱图。实际上，由于修饰的动态性，同一个肽的修饰和非修饰形式往往同时存在。这就给非限制性翻译后修饰检测提供了另一个线索——通过识别修饰-非修饰肽产生的相关谱图对来检测修饰。谱图网络算法就是基于这一原理，通过识别相关谱图对来检测翻译后修饰和突变等[84]。但是，谱图网络算法是利用 MS-alignment 算法计算谱图相似性，因而同样面临着计算量巨大的问题。另外，如果一个肽发生了某种修饰，但其非修饰形式不存在或者不能被质谱仪检测到，或者修饰-非修饰谱图对相似性不充分，则基于谱图对的方法就不适用了。

非限制性翻译后修饰检测作为蛋白质组学研究的最前沿，吸引了越来越多的研究者们进行尝试和探索。有研究者提出了先利用从头测试技术得到肽序列片段，再通过序列匹配定位蛋白质，进一步确定翻译后修饰质量和位点的方法[31, 85]。但是，这种策略严重依赖于质谱图的信号质量，而从头测序本身是个尚未很好解决的问题[33]。萨维茨基（Savitski）等人[86]提出了两种肽碎裂模式（ECD 和 CAD）联用的修饰检测方法，但是只适用于这种特殊的质谱操作模式。可以说，针对非限制性翻译后修饰检测的研究还在摇篮之中，目前尚没有成熟的解决方案。

6 计算所研制的 pFind 蛋白质鉴定系统

中科院计算所生物信息学研究组自 2002 年起开始研究基于生物质谱数据的蛋白质鉴定算法和软件，在质谱数据信号处理、理论质谱图预测、谱图相似性度量、蛋白质翻译后修饰检测、数据库索引以及搜索引擎设计等方面提出了一系列创新性算法和技术，在此基础上独立开发了我国第一个也是唯一一个蛋白质及其翻译后修饰的规模化鉴定软件系统 pFind (http://pfind.ict.ac.cn)。pFind 使用核谱向量点积（KSDP）作为核心匹配打分算法，对传统点积打分算法做了非线性扩展[58]；利用数据库索引、搜索流程和并行计算技术加速数据库搜索[87, 88]；利用谱图聚类方法快速地从质谱数据中检测潜在的修饰类型[89]。除核心搜索引擎外，pFind 还包括结果分析、谱图标注、数据库处理等多种配套支持软件[59, 60]，总计超过 21 万行代码。pFind 并行版也已经开始投入使用。目前，pFind 系统在精度和速度上已经达到国际主流商业软件如 SEQUEST 和 Mascot 的水平。pFind 系统已在国内外知名学术期刊和会议上发表学术论文 10 余篇[51, 58-60, 87-99]，并得到国际同行的认可和引用；申请发明专利 8 项，其中 3 项已获得授权；申请软件著作权 12 项。

pFind 系统目前已经在国内 10 余家蛋白质组研究单位示范应用，包括中科院上海生化细胞所、生物物理所、基因组所、动物所、化物所，北京蛋白质组研究中心、生命科学研究所、人类基因组北方中心、协和医科大学基础医学所和肿瘤所，以及上海生物信息中心、复旦大学等，总计安装 pFind 系统 62 套。2008 年，pFind 参加了 ABRF(生物分子资源实验室协会，Association of Biomolecular Resource Facilities)组织的国际蛋白质鉴定数据分析评测，在鉴定准确度和假阳性率控制能力方面表现出很强竞争力，开始在国际上崭露头角。北京蛋

白质组研究中心利用 pFind 系统鉴定核心岩藻糖修饰，从人类肝癌血浆样品中成功鉴定了100 余个核心岩藻糖修饰位点，是目前已有报道中最多的，对于后续癌症早期标记物发现研究意义重大。该合作成果于 2009 年发表在蛋白质组学领域国际著名期刊《分子与细胞蛋白组学（*Molecular & Cellular Proteomics*）》[100]，这是 pFind 软件第一次成功应用于真实生物学问题并得到国际一流学术期刊的认可。

7 总结

蛋白质组学研究方兴未艾，基于质谱数据的蛋白质及其修饰鉴定是其中最重要的问题之一。本文从数据库检索匹配打分、检索结果可靠性评估、修饰鉴定等几个方面介绍了蛋白质鉴定搜索引擎面临的关键计算问题。这些问题目前还没有良好的解决，有的已成为蛋白质组学数据分析的瓶颈。这一方面为计算领域提出了重大的挑战，另一方面也是计算技术发挥作用、解决生命科学问题的机会。我国在这方面的研究团队较少，但已经到达领域的最前沿。只要我们增强信心，加倍努力，pFind 蛋白质鉴定系统必将在未来的蛋白质组学研究中发挥更大的作用。

参考文献

[1] 钱小红, 贺福初: 蛋白质组学: 理论与方法. 北京: 科学出版社; 2003.

[2] Aebersold R, Mann M: Mass spectrometry-based proteomics. *Nature* 2003, 422:198-207.

[3] 杨芃原, 钱小红, 盛龙生: 生物质谱技术与方法. 北京: 科学出版社; 2003.

[4] 夏其昌, 曾嵘: 蛋白质化学与蛋白质组学. 北京: 科学出版社; 2004.

[5] Pappin DJ, Hojrup P, Bleasby AJ: Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993, 3(6):327-332.

[6] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20:3551-3567.

[7] Zhang W, Chait BT: ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* 2000, 72(11):2482-2489.

[8] Wilkins MR, Williams KL: Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J Theor Biol* 1997, 186(1):7-15.

[9] Clauser KR, Baker P, Burlingame AL: Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 1999, 71:2871-2882.

[10] Wells JM, McLuckey SA: Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* 2005, 402:148-185.

[11] Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 2004, 101(26):9528-9533.

[12] Sun S, Yu C, Qiao Y, Lin Y, Dong G, Liu C, Zhang J, Zhang Z, Cai J, Zhang H *et al*: Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *J Proteome Res* 2008, 7(1):202-208.

[13] Bafna V, Edwards N: SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001, 17:S13-S21.

[14] Eng JK, McCormack AL, Yates JR, III: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994, 5:976-989.

[15] Fenyö D, Qin J, Chait BT: Protein identification using mass spectromic information. *Electrophoresis* 1998, 19:998-1005.

[16] Field HI, Fenyo D, Beavis RC: RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* 2002, 2:36-47.

[17] Pevzner PA, Dancik V, Tang CL: Mutation-tolerant protein identification by mass-spectrometry. *J Comput Biol* 2000, 7:777-787.

[18] Zhang N, Aebersold R, Schwikowski B: ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2002, 2:1406-1412.

[19] Taylor JA, Johnson RS: Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* 1997, 11:1067-1075.

[20] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999, 6:327-342.

[21] Ma B, Zhang KZ, Hendrie C, Liang CZ, Li M, Doherty-Kirby A, Lajoie G: PEAKS: powerful software for peptide de novo sequencing by MS/MS. *Rapid Commun Mass Spectrom* 2003, 17:2337-2342.

[22] Frank A, Pevzner P: PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005, 77:964-973.

[23] Bern M, Goldberg D: EigenMS: de novo analysis of peptide tandem mass spectra by spectral graph partitioning. In: *Ninth Annual International Conference on Research in Computational Molecular Biology: 2005*; 2005: 357-372.

[24] Baginsky S, Cieliebak M, Gruissem W, Klemann T, Liptak Z, Muller M, Penna P: AUDENS: a tool for automated peptide de novo sequencing. *Journal of Proteome Research* 2005, 10:1768-1774.

[25] Chen T, Kao MY, Tepel M, Rush J, Church J: A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001, 8:325-337.

[26] Zhang Z: De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem* 2004, 76:6374-6383.

[27] Mann M, Wilm M: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994, 66:4390-4399.

[28] Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A: MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* 2003, 75:1307-1315.

[29] Frank A, Tanner S, Pevzner P: Peptide sequence tags for fast database search in mass-spectrometry. In: *Ninth Annual International Conference on Research in Computational Molecular Biology: 2005*; 2005: 326-341.

[30] Tabb DL, Saraf A, Yates JR, III: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003, 75:6415-6421.

[31] Han Y, Ma B, Zhang K: SPIDER: software for protein identification from sequence tags containing de Novo sequencing error. In: *IEEE 2004 Computational Systems Bioinformatics Conference: 2004*; 2004: 206-215.

[32] Johnson RS, Davis MT, Taylor JA, Patterson SD: Informatics for protein identification by mass spectrometry. *Methods* 2005, 35:223-236.

[33] Lu B, Chen T: Algorithms for de novo peptide sequencing via tandem mass spectrometry. *Drug Discovery Today: BioSilico* 2004, 2:85-90.

[34] Sadygov RG, Cociorva D, Yates JR, III: Large-scale database searching using tandem mass spectra:

looking up the answer in the back of the book. *Nat Methods* 2004, 1:195-202.

[35] Nesvizhskii AI, Vitek O, Aebersold R: Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007, 4(10):787-797.

[36] Wan KX, Vidavsky I, Gross ML: Comparing similar spectra: from similarity index to spectral contrast angle. *J Am Soc Mass Spectrom* 2002, 13:85-88.

[37] Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR, III: Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 2003, 75:2470-2477.

[38] Fenyö D, Beavis RC: A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 2003, 75:768-774.

[39] 盛泉虎, 汤海旭, 解涛, 王连水, 丁达夫: 用于串联质谱鉴定多肽的计量方法. *生物化学与生物物理学报* 2003, 35(8):734-740.

[40] Havilio M, Haddad Y, Smilansky Z: Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem* 2003, 75:435-444.

[41] Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RAJ, Speed TP, Simpson R.J: Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* 2003, 75:6251-6264.

[42] Sadygov RG, Yates JR, III: A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem* 2003, 75:3792-3798.

[43] Fridman T, Razumovskaya J, Verberkmoes N, Hurst G, Protopopescu V, Xu Y: The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *J Bioinform Comput Biol* 2005, 3:455-476.

[44] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: Open mass spectrometry search algorithm. *J Proteome Res* 2004, 3:958-964.

[45] Anderson DC, Li W, Payan DG, Noble WS: A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res* 2003, 2:137-146.

[46] Baczek T, Bucinski A, Ivanov AR, Kaliszan R: Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics. *Anal Chem* 2004, 76 (6):1726 -1732.

[47] Razumovskaya J, Olman V, Xu D, Uberbacher EC, VerBerkmoes NC, Hettich RL, Xu Y: A computational method for assessing peptide identification reliability in tandem mass spectrometry analysis with SEQUEST. *Proteomics* 2004, 4:961-969.

[48] Higdon R, Kolker N, Picone A, van Belle G, Kolker E: LIP index for peptide classification using MS/MS and SEQUEST search via logistic regression. *OMICS* 2004 8(4):357-369.

[49] Ulintz PJ, Zhu J, Qin ZS, Andrews PC: Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Molecular & Cellular Proteomics* 2006, 5:497-509.

[50] Liu J, Ma B, Li M: PRIMA: peptide robust identification from MS/MS spectra. In: *Third Asia-Pacific Bioinformatics Conference: 2005*; 2005: 181-190.

[51] Wang H, Fu Y, Sun R, He S, Zeng R, Gao W: An SVM Scorer for More Sensitive and Reliable Peptide Identification via Tandem Mass Spectrometry. In: *11th Pacific Symposium on Biocomputing: 2006*; 2006: 303-314.

[52] Washburn MP, Wolters D, Yates JR, III: Large-scale analysis of the yeast proteome via multidimensional protein identification technology. *Nat Biotech* 2001, 19:242-247.

[53] Li F, Sun W, Gao Y, Wang J: RScore: A Peptide Randomicity Score For Evaluating MS/MS Spectra.

Rapid Communications in Mass Spectrometry 2004, 18(14):1655-1659.

[54] Karlin S, Altschul S: Methods for assessing the statistical significance of molecular sequence features using general scoring schemes. *Proc Natl Acad Sci USA* 1990, 87:2264-2268.

[55] Karlin S, Altschul S: Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA* 1993, 90:5873-5877.

[56] Pearson WR: Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 1998, 276(1):71-84.

[57] Craig R, Beavis RC: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, 20:1466-1467.

[58] Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX, Gao W: Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* 2004, 20:1948-1954.

[59] Li D, Fu Y, Sun R, Ling C, Wei Y, Zhou H, Zeng R, Yang Q, He S, Gao W: pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* 2005, 21(13):3049-3050.

[60] Wang LH, Li DQ, Fu Y, Wang HP, Zhang JF, Yuan ZF, Sun RX, Zeng R, He SM, Gao W: pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2007, 21(18):2985-2991.

[61] Alves G, Ogurtsov AY, Yu YK: RAId_DbS: Peptide Identification using Database Searches with Realistic Statistics. *Biol Direct* 2007, 2:25.

[62] Kim S, Gupta N, Pevzner PA: Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* 2008, 7(8):3354-3363.

[63] Segal MR: On E-values for tandem MS scoring schemes. *Bioinformatics* 2008, 24(14):1652-1653; author reply 1654.

[64] Giddings JKaM: In response to ‘On E-value for tandem MS scoring schemes’. *Bioinformatics* 2008 24(14):1654.

[65] Alves G, Ogurtsov AY, Wu WW, Wang G, Shen RF, Yu YK: Calibrating E-values for MS2 database search methods. *Biol Direct* 2007, 2:26.

[66] Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 1995, 57(1):289-300.

[67] Keller A, Nesvizhskii AI, Kolker E, Aebersold R: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database Search. *Anal Chem* 2002, 74:5383-5392.

[68] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2003, 2(1):43-50.

[69] Elias JE, Gygi SP: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007, 4(3):207-214.

[70] Uy R, Wold F: Posttranslational covalent modification of proteins. *Science* 1977, 198(4320):890-896.

[71] Walsh CT: Posttranslational Modification of Proteins: Expanding Nature's Inventory. Englewood (Colorado): Roberts & Company Publishers; 2005.

[72] Nielsen ML, Savitski MM, Zubarev RA: Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics* 2006, 5(12):2384-2391.

[73] Hunyadi-Gulyás É, Medzihradszky KF: Factors that contribute to the complexity of protein digests. *Drug Discovery Today: TARGETS* 2004, 3(2, S1):3-10.

[74] Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A,

Williams KL *et al*: High-throughput mass spectrometric discovery of protein post-translational modifications. *J Mol Biol* 1999, 289(3):645-657.

[75] Mann M, Jensen ON: Proteomic analysis of post-translational modifications. *Nat Biotechnol* 2003, 21(3):255-261.

[76] Witze ES, Old WM, Resing KA, Ahn NG: Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 2007, 4(10):798-806.

[77] Creasy DM, Cottrell JS: Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2002, 2(10):1426-1434.

[78] Craig R, Beavis RC: A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* 2003, 17(20):2310-2316.

[79] Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G: Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* 2004, 4(9):2583-2593.

[80] Chamrad DC, Korting G, Schafer H, Stephan C, Thiele H, Apweiler R, Meyer HE, Marcus K, Bluggel M: Gaining knowledge from previously unexplained spectra-application of the PTM-Explorer software to detect PTM in HUPO BPP MS/MS data. *Proteomics* 2006, 6(18):5048-5058.

[81] Pevzner PA, Mulyukov Z, Dancik V, Tang CL: Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res* 2001, 11:290-299.

[82] Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA: Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 2005, 23(12):1562-1567.

[83] Chen Y, Chen W, Cobb MH, Zhao Y: PTMap--a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. *Proc Natl Acad Sci U S A* 2009, 106(3):761-766.

[84] Bandeira N, Tsur D, Frank A, Pevzner PA: Protein identification by spectral networks analysis. *Proc Natl Acad Sci U S A* 2007, 104(15):6140-6145.

[85] Na S, Jeong J, Park H, Lee KJ, Paek E: Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol Cell Proteomics* 2008, 7(12):2452-2463.

[86] Savitski MM, Nielsen ML, Zubarev RA: ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* 2006, 5(5):935-948.

[87] 87. Li D, Gao W, Ling CX, Wang X, Sun R, He S: IndexToolkit: an open source toolbox to index protein databases for high-throughput proteomics. *Bioinformatics* 2006, 22(20):2572-2573.

[88] Li Y, Chi H, Wang L-H, Wang H-P, Fu Y, Yuan Z-F, Li S-J, Liu Y-S, Sun R-X, Zeng R *et al*: Speeding up Tandem Mass Spectrometry Database Searching by Peptide and Spectrum Indexing. *Accepted by Rapid Communications in Mass Spectrometry* 2010.

[89] Fu Y, Jia W, Lu Z, Wang H, Yuan Z, Chi H, Li Y, Xiu L, Wang W, Liu C *et al*: Efficient discovery of abundant post-translational modifications and spectral pairs using peptide mass and retention time differences. *BMC Bioinformatics* 2009, 10 Suppl 1:S50.

[90] Fu Y, Sun R, Yang Q, He S, Wang C, Wang H, Shan S, Liu J, Gao W: A Block-Based Support Vector Machine Approach to the Protein Homology Prediction Task in KDD Cup 2004. *SIGKDD Explorations* 2004, 6:120-124.

[91] Fu Y, Yang Q, Ling CX, Wang H-P, Li D-Q, Sun R-X, Zhou H, Zeng R, Chen Y, He S-M *et al*: A Kernel-based Case Retrieval Algorithm with Application to Bioinformatics. *LNAI 3157* 2004:544-553.

[92] Zhang J, Gao W, Cai J, He S, Zeng R, Chen R: Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM T Comp Biol Bioinfo* 2005, 2(3):217-230.

[93] Zhang JF, He SM, Cai JJ, Cao XJ, Sun RX, Fu Y, Zeng R, Gao W: Preprocessing of tandem mass spectrometric data based on decision tree classification. *Genomics Proteomics Bioinformatics* 2005, 3(4):231-237.

[94] 王海鹏, 付岩, 孙瑞祥, 贺思敏, 曾嵘, 高文: pepReap:基于支持向量机的肽鉴定算法. *计算机研究与发展* 2005, 42(9):1511-1518.

[95] 孙瑞祥, 付岩, 李德泉, 张京芬, 王晓彪, 盛泉虎, 曾嵘, 陈益强, 贺思敏, 高文: 基于质谱技术的计算蛋白质组学研究. *中国科学 E 辑 信息科学* 2006, 36(2):222-234.

[96] Fu Y, Gao W, He S, Sun R, Zhou H, Zeng R: Mining tandem mass spectral data for more accurate mass error model for peptide identification. In: *12th Pacific Symposium on Biocomputing: 2007*; 2007: 421-432.

[97] Zhang J, He S, Ling CX, Cao X, Zeng R, Gao W: PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Commun Mass Spectrom* 2008, 22(8):1203-1212.

[98] Zhang J, Xu D, Gao W, Lin G, He S: Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun Mass Spectrom* 2009, 23(21):3448-3456.

[99] 孙瑞祥, 董梦秋, 迟浩, 杨兵, 秀丽蕴, 王乐珩, 付岩, 贺思敏: 基于电子捕获裂解/电子转运裂解串联质谱技术的蛋白质组学研究. *生物化学与生物物理进展* 2010, 37(1).

[100] Jia W, Lu Z, Fu Y, Wang HP, Wang LH, Chi H, Yuan ZF, Zheng ZB, Song LN, Han HH *et al*: A strategy for precise and large scale identification of core fucosylated glycoproteins. *Mol Cell Proteomics* 2009, 8(5):913-923.

作者简介：

付　岩　中国科学院计算技术研究所前瞻研究实验室，副研究员；yfu@ict.ac.cn

贺思敏　中国科学院计算技术研究所前瞻研究实验室，研究员；

孙瑞祥　中国科学院计算技术研究所前瞻研究实验室，副研究员；

王乐珩　中国科学院计算技术研究所前瞻研究实验室，工程师；